Co-Writing with Opinionated Language Models Affects Users' Views

Year: 2023 | Citations: 275 | Authors: Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, Mor Naaman

---

Abstract

If large language models like GPT-3 preferably produce a particular point of view, they may influence people's opinions on an unknown scale. This study investigates whether a language-model-powered writing assistant that generates some opinions more often than others impacts what users write – and what they think. In an online experiment, we asked participants (N=1,506) to write a post discussing whether social media is good for society. Treatment group participants used a language-model-powered writing assistant configured to argue that social media is good or bad for society. Participants then completed a social media attitude survey, and independent judges (N=500) evaluated the opinions expressed in their writing. Using the opinionated language model affected the opinions expressed in participants' writing and shifted their opinions in the subsequent attitude survey. We discuss the wider implications of our results and argue that the opinions built into AI language technologies need to be monitored and engineered more carefully.